MODELING THE PROSODY OF HIDDEN EVENTS FOR IMPROVED WORD RECOGNITION

Andreas Stolcke *Elizabeth Shriberg* *Dilek Hakkani-Tür* *Gökhan Tür*

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, U.S.A.
http://www.speech.sri.com/
{stolcke,ees,dilek,gokhan}@speech.sri.com

ABSTRACT

We investigate a new approach for using speech prosody as a knowledge source for speech recognition. The idea is to penalize word hypotheses that are inconsistent with prosodic features such as duration and pitch. To model the interaction between words and prosody we modify the language model to represent hidden events such as sentence boundaries and various forms of disfluency, and combine with it decision trees that predict such events from prosodic features. N-best rescoring experiments on the Switchboard corpus show a small but consistent reduction of word error as a result of this modeling. We conclude with a preliminary analysis of the types of errors that are corrected by the prosodically informed model.

1. INTRODUCTION

One source of information that currently is not being explicitly modeled for large-vocabulary speech recognition is prosody: the suprasegmental duration, pitch, and energy features of speech. Prosodic cues have been used in automatic speech processing systems for various tasks, such as lexical and syntactic disambiguation, dialog processing and speech understanding [11, 7, among others]. Research on small- and medium-vocabulary recognizers have shown that prosodic cues can raise the rank of the correct hypothesis [17, 16]. However, prosody is currently not widely used in large-vocabulary word recognition.

One difficulty in leveraging prosody for word recognition is that it correlates with linguistic structures that are mainly at or above the word level; models based on local likelihoods (similar to the standard acoustic models of today's recognizers) are unsuitable. Therefore, it seems more promising to leverage prosody for word recognition in an indirect way: we model the higher-level structures that manifest themselves prosodically, as well as the relationship between these structures and the word sequence, and evaluate a word hypothesis based on the consistency of all three components: words, structure, and prosody [16]. For example, we might have a model of syntactic structure and its prosodic manifestations, as well as a word language model in terms of syntactic structure. Both together can be used to penalize hypotheses whose likely syntactic structure is inconsistent with prosody, and to boost those that are consistent with it.

In this paper, we instantiate this idea, using linguistic structure of a more rudimentary kind. Instead of full-fledged syntax, we model the prosody and word sequences associated with sentence boundaries and certain types of disfluencies (hesitations and self-repairs). We refer to both types of phenomena as *hidden events*, because they can be thought of as hidden pseudo-words occurring between the observable words. For example,

> Right <S> I <REP> I don't <DEL> uh <FP> I'm not really sure . . .

shows a sentences boundary <S>, a disfluent repetition <REP>, and a disfluent deletion (false start) <DEL> as tags at their respective locations in the word stream. These are the kinds of events we will model, both prosodically and lexically.

In the next section we formalize the general approach to leveraging prosody via linguistic structure. Section 3 elaborates on the modeling of hidden events and how they can be fit into the framework. Section 4 presents some preliminary experiments, analyses, and examples of corrected recognition errors. Section 5 discusses further work and Section 6 concludes.

2. MODELING APPROACH

Before going into the specifics of hidden-event modeling, we can formulate the approach outlined in the Introduction in formal terms. We will denote word sequences with W, the associated standard acoustic features with A, and any prosodic features with F. Given an acoustic manifestation, a standard speech recognizer searches for the word sequence with highest posterior probability, which can be estimated using a word language model $P(W)$ and an acoustic likelihood model $P(A|W)$ [1]:

$$\begin{aligned} W^* &= \operatorname*{argmax}_W P(W|A) \\ &= \operatorname*{argmax}_W \frac{P(W)P(A|W)}{P(A)} \\ &= \operatorname*{argmax}_W P(W)P(A|W) \end{aligned} \tag{1}$$

Now let us assume that, in order to leverage prosody, we ascribe a structure S to the word sequence W. S could be a parse tree, or, in our case, a representation of the hidden events (sentence boundaries, disfluencies) embedded in W. We also assume that we have a model for the relation between words, prosody, and structure, i.e., $P(W, S, F)$. Again, the motivation for S is that it is easier to model $P(W, S, F)$ than a direct relation $P(W, F)$ between words and prosody. The details of this model are unimportant for now. We can revise Equation 1 to condition the word hypotheses on both the standard acoustic features A and the prosodic features F:

$$\begin{aligned} W^* &= \operatorname*{argmax}_W P(W|A, F) \\ &= \operatorname*{argmax}_W \frac{P(W|F)P(A|W, F)}{P(A|F)} \\ &\approx \operatorname*{argmax}_W \frac{P(W|F)P(A|W)}{P(A|F)} \\ &= \operatorname*{argmax}_W P(W|F)P(A|W) \\ &= \operatorname*{argmax}_W P(W, F)P(A|W) \\ &= \operatorname*{argmax}_W \sum_S P(W, S, F)P(A|W) \end{aligned} \tag{2}$$

Line 3 relies on the approximation that the standard acoustic features are independent of the prosody once conditioned on the word sequence. For line 4, we use the fact that $P(A|F)$ and $P(F)$ are constants with respect to W. Note that the last line requires us to consider all possible structures S for a given word sequence.

Event class	Tag	Freq.	Example
Sentence boundary	**S**	10.8%	I haven't seen it $*$ Not sure I like it
Filled pause	**FP**	2.9%	he uh $*$ liked it
Repetition	**REP**	1.9%	he $*$ he liked it
Deletion	**DEL**	1.3%	it was $*$ he liked it
Repair	**OthDF**	1.2%	he $*$ she liked it
Else/fluent	**else**	81.8%	she $*$ liked it

Table 1. Boundary and disfluency event classes.

3. HIDDEN EVENTS

The present work builds on our previous research on modeling hidden events for the purpose of automatic detection [12, 15]. Hidden events can be viewed as tags that label the type of boundary between adjacent words. We used the sentence boundary and disfluency event classes from [15] in our models, shown in Table 1 with examples and frequencies in the corpus we used for experiments.

3.1. Prior Work

The hidden-event classes chosen correlate with the surrounding words, as well as with prosodic features such as pause, duration, and pitch. Hidden events are thus suitable candidates for the kind of hidden structure needed to leverage prosody as a knowledge source for word recognition. Prosodic cues have been studied mainly for the purpose of automatic detection of disfluencies [10, 12] and sentence boundaries [8]. The correlation between hidden events and word cues has likewise been exploited, for detecting both sentence boundaries [14, 8] and disfluencies [2, 6, among others], although recent work has also shown that speech language models can be improved by incorporating hidden events into the model [5]. For the present work we reused prosodic and language models of hidden events previously developed for automatic detection from combined acoustic and lexical cues [15], but applying the models in the word recognition paradigm of Section 2.

Compared to other work on word recognition, our approach is most similar to the prosody/parse scoring paradigm of Veilleux and Ostendorf [16], who also propose leveraging prosody for word recognition through hidden structure, in a probabilistic framework. In their case, the hidden structure consists of the syntactic parse of the utterance. Another difference is that we model continuous prosodic features directly from the hidden structure, rather than using an intermediate phonological representation (prominence labels and break indices).

3.2. Hidden-Event Modeling

Our goal is to model the joint probabilities $P(W, S, F)$ of words W, hidden structure S, and prosodic features F. The hidden structure in this case consists of a sequence of events $S = E_1, E_2, \ldots E_n$, corresponding to the words boundaries following the words $W = W_1, W_2, \ldots, W_n$. The E_i are from the set shown in Table 1.

We decompose $P(W, S, F)$ into the joint probability of words and events, and that of the prosody given the words and events:

$$P(W, S, F) = P(W, S)P(F|W, S) \tag{3}$$

Furthermore, we assume that the prosodic features correlate with the events in a local fashion: prosodic features F_i are computed from a window around boundary i, and correlate mainly with event E_i:

$$\begin{aligned} P(F|W, S) &= P(F_1 \ldots F_n|E_1 \ldots E_n, W) \\ &\approx \prod_{i=1}^{n} P(F_i|E_i, W) \end{aligned} \tag{4}$$

For modeling the relation between words and events, $P(W, S)$, we use standard language modeling techniques. The events can be represented as pseudo-words and the whole sequence $(W, S) = W_1 E_1 W_2 E_2 \ldots W_n E_n$ may be modeled using a standard N-gram model. The model is trained on annotated transcripts using standard smoothing and backoff techniques. To make better use of the limited span of the N-gram model, we represented only sentence boundary and disfluency events by tags; sentence-internal fluent word transition events (which account for the vast majority of cases; cf. Table 1) are represented implicitly by the absence of an event tag, as shown in the example from the Introduction:

> Right <S> I <REP> I don't <DEL> uh <FP> I'm not really sure . . .

During testing, the events are unknown. According to Equation 2, we need to sum over all possible event sequences for a given word sequence. By using an N-gram model for $P(W, S)$, and decomposing the prosodic likelihoods as in Equation 4, the joint model $P(W, S, F)$ becomes equivalent to a hidden Markov model (HMM). The HMM states are the (word,event) pairs, while prosodic features form the observations. Transition probabilities are given by the N-gram model; emission probabilities are estimated by the prosodic model described below. Based on this construction, we can carry out the summation over all possible event sequences efficiently with the familiar forward dynamic programming algorithm for HMMs.

3.3. Prosodic Model

We are thus left with the task of estimating likelihoods of events E_i, $P(F_i|E_i, W)$, based on prosodic features F_i around a word boundary. Because the event space is discrete and small, and the prosodic feature space continuous, high-dimensional, and highly correlated, it is convenient to invert the problem and model posterior event probabilities instead:

$$P(F_i|E_i, W) = \frac{P(F_i|W)P(E_i|F_i, W)}{P(E_i|W)}$$

We assume that the prosodic features are marginally independent of the words, $P(F_i|W) \approx P(F_i)$, so that the first term can be treated as a constant. This is justified if, as in our case, we only make prosodic features dependent on segmentation information from an alignment of W, and not on the identities of the words themselves.

The posterior event probabilities $P(E_i|F_i, W)$ could be estimated by a variety of probabilistic classifiers, including decision trees, neural networks, or exponential models. As in previous work, we trained CART-style decision trees [3] to estimate these posteriors, and resampled the data to give equal priors for all event types, so as to avoid the explicit scaling by $P(E_i|W)$. (Note that this effectively scales by $P(E_i)$ only, again justified by the weak conditioning on alignment information.)

4. RESULTS AND DISCUSSION

We tested our approach on the Switchboard corpus of conversational speech [4]. Prosodic and event language models were trained on 900 conversations that had been annotated with hidden events by the Linguistic Data Consortium [9]. We generated the top 100 hypotheses for a 19-conversation test set (18,000 words), using standard acoustic and language models.

A further six conversations were decoded for the purpose of tuning model parameters. Two weighting parameters were optimized this way: first, an exponent on $P(W, F)$ in Equation 2 serves to balance the overall prosodically-conditioned model with the standard acoustic model. This parameter corresponds to the language model weight in a standard recognizer. Second, an exponent on $P(F|W, S)$ in Equation 3 balances the influence of the prosodic component against the event language model $P(W, S)$.

The N-best lists were rescored with three models:

Model	WER (%)	Sub	Del	Ins
Standard N-gram	47.9	31.1	12.2	4.6
HE N-gram, no prosody	47.6	30.4	13.3	3.9
HE N-gram, with prosody	47.0	29.7	14.1	3.2

Table 2. Results rescoring 100-best lists with hidden-event models. The word error rates (WER) is broken down into Sub(stitutions), Del(etions) and Ins(ertions).

- A standard trigram, trained on the same amount of data as the hidden-event models (a 4-gram model was generated but proved no better than the trigram).
- A hidden-event 4-gram model (without prosodic conditioning). (This model is obtained by setting the second tuning parameter to zero.)
- A prosodically conditioned 4-gram hidden-event model (the full model proposed in this paper).

All weights (including the LM weight for the standard model) were tuned independently on the tuning set.

The prosodic features used captured a range of durational aspects of the speech. They included the duration of pauses, of final vowels and of final syllable rhymes, which were normalized both for phone duration and by speaker-specific statistics. Notably, no features based on pitch and energy were used, as such features had not proven helpful for event detection in past work on this corpus.

4.1. Word Error Results

Table 2 summarizes the results. We see a small (0.9% absolute reduction) in word error rate (WER) between the baseline and the full, prosodically conditioned hidden-event model. (The difference is highly significant in a matched-pairs test, $p < .000001$.)

Note that about one third of the improvement seems to come from hidden-event modeling in the language model alone (0.3%, $p < .02$). This is evidence that, even without prosodic conditioning, the hidden-event language model does a better job at modeling the words, by considering the potential events between words.

4.2. Result Analysis

The breakdown of the WER by error type in Table 2 shows that the overall reduction in error is achieved through fewer substitutions and insertions, at the expense of more deletions. Note that this is not due to a differently optimized word insertion penalty (the word insertion penalty was fixed at 0, a value that happened to be optimal for the baseline model). Thus, the hidden-event model seems to inherently suppress insertions and substitutions. This trend is present even for the word-only hidden-event model. and is further reinforced by the use of prosodic cues.

A preliminary error analysis suggests that the prosodic model reduces false detections of high-frequency words that tend to occur at sentence boundaries and/or in disfluent repetitions ("I", "and", "the"). Since such words are very frequent in the training data, and are often phonologically reduced, they are likely candidates for misrecognitions; our model has the means to suppress them except in cases where the prosody is consistent with sentence boundaries and/or disfluencies.

In order to better understand *how* the event modeling provided the win in word accuracy, we conducted a high-level diagnostic analysis comparing the baseline model to the prosodic hidden-event model. We used the following approach to sort output into useful subsets for analysis. For each word in the reference transcript, we aligned recognized word strings for each of the three models. Insertions were arbitrarily grouped as errors with the following word; while this is suboptimal (using temporal or phonetic distances to determine attachment would be clearly preferable) it does not change overall error counts.

Thus for each reference word we obtained an error type associated with the model of interest, where the error type could be CORRECT (no error), DELETION, SUBSTITUTION, or a combination of (one or more) INSERTION with CORRECT or SUBSTITUTION. Results could thus be compared across models, indexed by reference word.

Baseline	HE model	Ref. words	Error Δ
correct	incorrect	11402	0
incorrect	incorrect	8042	-118
correct	incorrect	597	+606
incorrect	correct	569	-619

Table 3. Breakdown of changes in error status of reference words between baseline and prosodic hidden-event models.

Overall statistics from this analysis are tabulated in Table 3. For example, we see that 8042 reference words were incorrectly recognized (or preceded by insertions) in both baseline and HE model. However, due to insertions, the total number of errors attributed to these cases was 118 less in the HE model. Similarly, although slightly more reference words were incorrectly recognized (or attached to insertions) in the HE model than in the baseline, the resulting word error count still comes out in favor of the HE model.

4.3. Examples

More detailed analyses are needed to better characterize the nature of errors corrected by the hidden-event model. However, using the analysis described above, we were able to extract prototypical cases that illustrate how the hidden-event model can correct errors. The examples below are taken from the set of errors that were present even with the word-only hidden-event model, and then corrected in the prosody-informed model; we can therefore attribute these corrections to the influence of prosody.

Right before a sentence boundary event, the baseline model allows word sequences to end in a sentence fragment, but the hidden-event model strongly disprefers words or N-grams that cannot end a sentence. The example below illustrates how the phonetically similar "to", a more frequent word than the correct "too", is fit for a sentence fragment "at church to. . . ", but not for the sentence-final "at church to.", thus giving preference to "too" in this context.

(2131-B-0053) . . . that at church to <S>
⟶ . . . that at church too <S>

Filled pauses are frequent disfluencies, whose prosodic features (particularly duration, but also surrounding pauses) are useful for discriminating them from other frequent short words:

(3528-B-0038) . . . to perform in and col weather
⟶ . . . to perform in UH cold weather

Repetitions are another common disfluency which have characteristic prosodic patterns that distinguish them from both fluent repeats and from fluent nonrepeated sequences [13]. While disfluent repetitions like "the the" are represented in a non-event language model, they are less frequent than other, fluent sequences like "to the". Thus actual repetitions are often misrecognized as non-repeated sequences, as in

(2461-B-0044) . . . to really hurt to <REP> the middle class
⟶ . . . to really hurt the <REP> the middle class

The event modeling also worked in the other direction, e.g., preventing spurious disfluent sequences when the speech has fluent prosody. For instance, as just mentioned, repeated words from disfluent repetitions such as "the the" are implicitly represented in a non-event language model. However, the prosodic event model can prevent such cases from surfacing in recognition if the characteristic repetition prosody is not detected in the context, as it did in the following example:

(2753-A-0008) . . . problem is here <S> the the source of . . .
⟶ . . . problem is here <S> but the source of . . .

This effect generalized to non-event regions as a whole. That is, while we might expect that the effects of event modeling are greatest for words bordering on events, in actuality about 70% of cases in which there was an error in the baseline model and a correct hypothesis in the event model, were in non-event contexts.

5. ISSUES FOR FUTURE WORK

We do not want to claim that hidden events are a particularly effective way to bring prosodic information to bear on word recognition. In fact, since sentences boundaries and disfluencies together occur at only 18% of word boundaries, the effect of the model tried here is inherently limited. The experiments serve mainly as a proof of concept, while many other aspects of prosody remain to be exploited.

It should be noted that our comparison to the baseline model underestimates the importance of prosodic cues for word recognition, as the baseline model already uses prosodic cues implicitly. This is due to the common practice in conversational speech recognition to chop waveforms at long pauses prior to recognition; the language model is conditioned on the location of these chopping boundaries, and hence on pauses.

The hidden-event modeling approach itself is not fully explored. For example, we have not yet tried to optimize the feature set used by the prosodic model for word recognition. For convenience we used an existing prosodic decision tree optimized for *event* (rather than word) recognition. The tree used only durational features; pitch and energy features might yield additional improvements.

Another avenue for further exploration involves varying the set of events modeled. It is possible that focusing on the high-frequency events (sentence boundaries, filled pauses and repetitions) is better for word recognition. On the other hand we could include events such as discourse markers, which are prosodically distinct and have been shown to be beneficial to language and boundary modeling [5].

6. CONCLUSION

We have argued for an indirect way to incorporate prosody into word recognition by modeling hidden structure, beyond the words, that correlates with prosody. A simple example of this approach is the modeling of hidden events (sentence boundary, disfluencies) in spontaneous speech, since such events are partly marked by their prosodic manifestations, and are also correlated with specific word choices. With some independence assumptions, hidden events can be modeled efficiently by a combination of hidden Markov model and prosodic decision trees. Experiments show that this type of model can reduce the word error rate of a large-vocabulary recognizer. The improvement is obtained by boosting word hypotheses that are consistent with hidden-event prosody. Empirically, errors are corrected in both event and non-event contexts, by suppressing insertion errors and false detections of disfluent word sequences.

ACKNOWLEDGMENTS

We thank Becky Bates and Mari Ostendorf for assistance in preparing the data and valuable discussions. This research was in part supported by DARPA and NSF under NSF grant IRI-9619921. The views herein are those of the authors and should not be interpreted as representing the policies of the funding agencies.

REFERENCES

[1] L. R. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *IEEE PAMI*, 5(2):179–190, 1983.

[2] J. Bear, J. Dowding, and E. Shriberg. Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog. In *Proc. ACL*, pp. 56–63, University of Delaware, Newark, Delaware, 1992.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Pacific Grove, CA, 1984.

[4] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. ICASSP*, vol. 1, pp. 517–520, San Francisco, 1992.

[5] P. Heeman and J. Allen. Intonational boundaries, speech repairs, and discourse markers: Modeling spoken dialog. In *Proc. ACL/EACL*, Madrid, 1997.

[6] P. A. Heeman and J. Allen. Detecting and correcting speech repairs. In *Proc. ACL*, pp. 295–302, New Mexico State University, Las Cruces, NM, 1994.

[7] R. Kompe. *Prosody in speech understanding systems*. Springer, Berlin, 1997.

[8] M. Mast, R. Kompe, S. Harbeck, A. Kießling, H. Niemann, E. Nöth, E. G. Schukat-Talamazzini, and V. Warnke. Dialog act classification with the help of prosody. In H. T. Bunnell and W. Idsardi, editors, *Proc. ICSLP*, vol. 3, pp. 1732–1735, Philadelphia, 1996.

[9] M. Meteer, A. Taylor, R. MacIntyre, and R. Iyer. Dysfluency annotation stylebook for the Switchboard corpus. Distributed by LDC, `ftp://ftp.cis.upenn.edu-/pub/treebank/swbd/doc/DFL-book.ps.gz`, 1995. Revised June 1995 by Ann Taylor.

[10] C. H. Nakatani and J. Hirschberg. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America*, 95(3):1603–1616, 1994.

[11] M. Ostendorf, C. Wightman, and N. Veilleux. Parse scoring with prosodic information: an analysis/synthesis approach. *Computer Speech and Language*, 4:193–210, 1993.

[12] E. Shriberg, R. Bates, and A. Stolcke. A prosody-only decision-tree model for disfluency detection. In G. Kokkinakis, N. Fakotakis, and E. Dermatas, editors, *Proc. EUROSPEECH*, vol. 5, pp. 2383–2386, Rhodes, Greece, 1997.

[13] E. E. Shriberg. Acoustic properties of disfluent repetitions. In *Proceedings International Congress of Phonetic Sciences*, vol. 4, pp. 384–387, Stockholm, 1995.

[14] A. Stolcke and E. Shriberg. Automatic linguistic segmentation of conversational speech. In H. T. Bunnell and W. Idsardi, editors, *Proc. ICSLP*, vol. 2, pp. 1005–1008, Philadelphia, 1996.

[15] A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tür, and Y. Lu. Automatic detection of sentence boundaries and disfluencies based on recognized words. In R. H. Mannell and J. Robert-Ribes, editors, *Proc. ICSLP*, vol. 5, pp. 2247–2250, Sydney, 1998. Australian Speech Science and Technology Association.

[16] N. M. Veilleux and M. Ostendorf. Prosody/parse scoring and its applications in ATIS. In *Proc. ARPA HLT Workshop*, pp. 335–340, Plainsboro, NJ, 1993.

[17] A. Waibel. Prosodic knowledge sources for word hypothesization in a continuous speech recognition system. In *Proc. ICASSP*, pp. 20.16.1–20.16.4, Dallas, Texas, 1987.